Learning and Optimal Control of Imprecise Markov Decision Processes by Dynamic Programming Using the Imprecise Dirichlet Model

Matthias C. M. Troffaes

Ghent University, SYSTeMS Research Group, Technologiepark Zwijnaarde 914, 9052 Zwijnaarde, Belgium. Matthias.Troffaes@UGent.be

Summary. In this paper, we investigate the conditions under which dynamic programming yields a solution to simultaneous learning and optimal control of a Markov decision process. First, we introduce a new optimality criterion that allows act-state dependence. This criterion is based on a partial preference ordering induced by an imprecise probability model of the dynamics of the system, updated by observations of the state and control history of the system. Then, we show that dynamic programming yields the set of all optimal solutions if the imprecise probability model satisfies particular properties. When we model learning of the system dynamics by an imprecise Dirichlet model, these properties turn out to be satisfied.

1 Introduction

Already early in the development of the theory of Markov decision processes, it was recognised that the transition probabilities themselves are often subject to uncertainty, simply because they are often hard to measure. Two solutions have been suggested and studied in the literature: (i) *learning*—update the transition probabilities as we observe transitions [4, 5], and (ii) *sets*—only assume the transition probabilities belong to some convex set [5, 7, 2, 3].

Unfortunately, the learning-based solution relies heavily on prior information about the transition probabilities. If this prior information is incorrect, the optimal policy can be subject to serious bias in the initial phase of the process. A drawback of the set-based solution is that it does not involve learning, ignoring possibly useful information that is often available. Moreover, it has a problematic relation with optimality: working with a set of transition probabilities, we can only associate an *interval* for the expected reward with each control law. Most authors therefore have considered only extreme solutions, developing algorithms to find control laws that either maximise the minimal expected gain (pessimistic), or that maximise the maximal expected gain (optimistic), ignoring solutions not relying on such extreme assumptions.

One notable exception is in [3], where an algorithm is suggested to find the set of all maximal elements with respect to a partial preference order, based on comparing intervals. In that way, not only the extreme solutions are recovered. However, in [3] it is not questioned in what sense the proposed dynamic programming method leads to optimal policies. In this article we approach the problem from a more logical side: we *first* define a notion of optimality and investigate whether the dynamic programming argument holds for this notion of optimality, instead of blindly "generalising" Bellman's algorithm.

We have previously shown in [1] that dynamic programming works if and only if our notion of optimality satisfies two conditions: (i) the *principle of optimality*, and (ii) *insensitivity with respect to omission of non-optimal elements*. Unfortunately, the first condition does not hold when using the partial order of [3] (see counterexample in [1]). Hence, the algorithm of [3] actually does *not* result in *optimal* control laws in the sense of maximality with respect to the suggested order. In [1], a different partial order is suggested for deterministic systems with uncertain gain, which does satisfy the principle of optimality and the insensitivity property. However, this order does not simply generalise to non-deterministic systems. The reason is act-state dependence.

Our goal is combining the learning-based solution with the set-based solution, overcoming the problems from which each method separately suffers. Basically, we update the set of transition probabilities based on observations of previous transitions, *e.g.*, through an imprecise Dirichlet model. First, we generalise the orders used in [1, 3] to the case of act-state dependence. Then we show that there are fairly general conditions under which the principle of optimality and the insensitivity property still hold.

Section 2 introduces some aspects of imprecise probabilities [6] used further on. Section 3 motivates a new partial preference order allowing act-state dependence. Section 4 defines the systems studied and describes how to compare control laws. Section 5 states conditions for the principle of optimality to hold, and considers simultaneous learning and control of a Markov decision process by an imprecise Dirichlet model. Section 6 concludes the paper.

2 Lower Previsions and Marginal Extension

Let X be a random variable (such as the state of a system at a particular time) taking values in a set $\mathcal{X}$. A particular value of X is denoted by x. A *gamble* f on X is a bounded $\mathcal{X}$–$\mathbb{R}$ map. It is an uncertain reward: if x turns out to be the true value of X, we receive an amount $f(x)$ of utility. $\mathcal{L}(X)$ denotes the set of all gambles on X. We may write $f(X)$ to emphasise that f is a gamble on X. We define the gamble $I_{X=x'}$ as $I_{X=x'}(x) = \begin{cases} 1, & x=x', \\ 0, & x\neq x'. \end{cases}$

Define the *lower prevision* $\underline{P}(f)$ of the gamble f as the supremum buying price for f: for any $s < \underline{P}(f)$, we are willing to pay s prior to observation of X, if we are guaranteed to receive $f(x)$ when x turns out to be the value of X. We can also interpret f as an uncertain loss: if x turns out to be the true

value of X, we lose $f(x)$. The *upper prevision* $\overline{P}(f)$ of the gamble f is the infimum selling price for f: for any $s > \overline{P}(f)$, we are willing to receive s prior to observation of X, if we are guaranteed to lose $f(x)$ when x turns out to be the value of X. Since a reward r is a loss $-r$ it holds that $\overline{P}(f) = -\underline{P}(-f)$.

If $\underline{P}$ is defined on all gambles in $\mathcal{L}(X)$, it should satisfy for all $f, g \in \mathcal{L}(X)$:

- $\underline{P}(f) \geq \inf_{x \in \mathcal{X}}[f(x)]$ (accepting sure gain)
- $\underline{P}(\lambda f) = \lambda \underline{P}(f)$, whenever $\lambda \in \mathbb{R}$ and $\lambda > 0$ (scale independence)
- $\underline{P}(f+g) \geq \underline{P}(f) + \underline{P}(g)$ (price of sum at least sum of prices of each term)

In such a case we call $\underline{P}$ *coherent*. If it also holds that

- $\underline{P}(f+g) = \underline{P}(f) + \underline{P}(g)$ (price of sum *equal to* sum of prices of each term)

then we call $\underline{P}$ *linear*. Linear lower previsions are expectations in the sense of classical probability theory, and satisfy $\underline{P}(f) = \overline{P}(f)$ for all $f \in \mathcal{L}(X)$. Therefore the bars are usually dropped and P is simply called a *linear prevision*.

Let us now consider two variables, say X and Y. The supremum buying price of a gamble f on X conditional on the value y of Y is denoted by $\underline{P}(f|y)$. Through *separate coherence*, the coherent conditional lower previsions $\underline{P}(\cdot|y)$ on $\mathcal{L}(X)$, defined for all $y \in \mathcal{Y}$, jointly extend to a $\mathcal{L}(X,Y)$–$\mathcal{L}(Y)$-map $\underline{P}(\cdot|Y)$:

$$\underline{P}(f(X,Y)|Y)(y) := \underline{P}(f(X,y)|y), \tag{1}$$

for any gamble $f(X,Y)$. Here, $f(X,y)$ denotes a gamble on X by fixing the value y of Y in f, *i.e.*, $f(X,y)(x) := f(x,y)$.

In case of n variables $X_1, \ldots, X_n$, conditional lower previsions $\underline{P}(\cdot|x_1)$ on $\mathcal{L}(X_2)$, $\underline{P}(\cdot|x_1x_2)$ on $\mathcal{L}(X_3)$, $\ldots$, and $\underline{P}(\cdot|x_1 \ldots x_{n-1})$ on $\mathcal{L}(X_n)$ extend through separate coherence to a $\mathcal{L}(X_1,X_2)$–$\mathcal{L}(X_1)$-map, a $\mathcal{L}(X_1,X_2,X_3)$–$\mathcal{L}(X_1,X_2)$-map, $\ldots$, and a $\mathcal{L}(X_1,\ldots,X_n)$–$\mathcal{L}(X_1,\ldots,X_{n-1})$-map. Concatenating these, we end up with a $\mathcal{L}(X_1,\ldots,X_n)$–$\mathcal{L}(X_1)$-map, which is in fact a coherent lower prevision on all variables, conditional on X_1:

$$\begin{aligned}&\underline{P}(f(X_1,X_2,\ldots,X_n)|X_1)\\ &= \underline{P}(\cdot|X_1) \circ \underline{P}(\cdot|X_1X_2) \circ \cdots \circ \underline{P}(\cdot|X_1X_2\ldots X_{n-1})(f(X_1,X_2,\ldots,X_n))\end{aligned} \tag{2}$$

This lower prevision is the *marginal extension* of $\underline{P}(\cdot|X_1)$, $\underline{P}(\cdot|X_1X_2)$, $\ldots$, and $\underline{P}(\cdot|X_1\ldots X_{n-1})$. In the classical theory of probability (2) is Bayes rule.

3 Optimality in Case of Partial Act-State Dependence

We have a set of actions A and want to characterise the optimal actions. One way to do this is through a preference order on the actions, selecting as optimal the set of actions that are maximal with respect to this partial order:

Definition 1. *Let $>$ be a strict partial order on A. Then $a^* \in A$ is said to be* maximal *with respect to $>$ if there is no $a \in A$ such that $a > a^*$.*

Assume that $X = (\Xi, \Theta)$, and acts $a \in A$ do not influence the value of Θ. Thus, we model our knowledge about Θ by a coherent lower prevision $\underline{P}$ on $\mathcal{L}(\Theta)$, independent of the action a we take. Now assume that the act dependent information is modelled through a coherent conditional lower prevision $\underline{P}_a(\cdot|\theta)$ on $\mathcal{L}(\Xi)$, for each action $a \in A$ and each value θ of Θ. If

$$\underline{P}(\underline{P}_a(f_a|\Theta) - \overline{P}_b(f_b|\Theta)) > 0. \tag{3}$$

then we should prefer action a over action b. Indeed, by (3) we are willing to pay a strictly positive price prior to observation of Θ in order to receive $\underline{P}_a(f_a|\theta)$ and to lose $\overline{P}_b(f_b|\theta)$, if θ turns out to be the value of Θ, independently of the action we take. Hence, for some $\epsilon > 0$, we are, prior to observing Θ, willing to pay a strictly positive price in order to receive $\underline{P}_a(f_a|\theta) - \epsilon$ and to lose $\overline{P}_b(f_b|\theta) + \epsilon$ if θ has been observed. Suppose now θ has been observed. Then, using the behavioural interpretation of $\underline{P}_a(f_a|\theta)$, for any $\epsilon > 0$ we are willing to lose $\underline{P}_a(f_a|\theta) - \epsilon$ prior to observation of Ξ, in order to take action a and receive $f_a(\xi, \theta)$ after observation of $\Xi = \xi$. But, we are also willing to take action b and lose $f_b(\xi', \theta)$ after observation of $\Xi = \xi'$, if we receive $\overline{P}_b(f_b|\theta) + \epsilon$ prior to observation of Ξ. Combining all, prior to any observation of Ξ and Θ we are willing to pay a strictly positive price in order to exchange action b for a along with their rewards $f_b(\xi, \theta)$ and $f_a(\xi', \theta)$.

For example, if there is no partial act-state independence, we can identify Ξ with X, recovering the order used by [3]: $\underline{P}_a(f_a) > \overline{P}_b(f_b)$. On the other hand, in case of full act-state independence, we can identify Θ with X, recovering the order discussed in [6, Sect. 3.9] and used in [1]: $\underline{P}(f_a - f_b) > 0$.

4 Imprecise Statistical Decision Processes

Let $\mathcal{X}$ be the finite set of *states* the system can assume, and $\mathcal{U}$ the finite set of *controls* we can apply. The system state at time k is denoted by X_k, and x_k denotes a particular value of X_k. We are not interested in dynamics of the system beyond time N. Consider the system at time k and imagine

- observing $X_k = x_k$,
- applying $\mu_k(x_k) \in \mathcal{U}$ and observing $X_{k+1} = x_{k+1}$,
- applying $\mu_{k+1}(x_k x_{k+1}) \in \mathcal{U}$ and observing $X_{k+2} = x_{k+2}$,
- *etc.*,
- applying $\mu_{N-1}(x_k x_{k+1} \dots x_{N-1}) \in \mathcal{U}$ and observing $X_N = x_N$.

This operation is characterised by a finite sequence of functions $\pi_k = (\mu_k, \mu_{k+1}, \dots, \mu_{N-1})$, where $\mu_\ell \colon \mathcal{X}^{\ell-k+1} \to \mathcal{U}$. We call π_k a *control law* from time k, and Π_k denotes the set of all control laws from time k. For each control law π_k we have a *gain gamble from time ℓ after observation of $x_k \dots x_{\ell-1}$*,

$$J_{\pi_k(x_k \dots x_{\ell-1})}(x_\ell, \dots, x_N) = \textstyle\sum_{q=\ell}^{N-1} g_q(x_q, \mu_q(x_k \dots x_q), x_{q+1}) + g_N(x_N) \tag{4}$$

It is a gamble on $(X_\ell, \ldots, X_N)$. Each transition incurs a gain: starting at time q in x_q, applying u_q and arriving in x_{q+1}, we receive an amount $g_q(x_q, u_q, x_{q+1})$. Arriving in the final state x_N at time N, we receive an additional gain $g_N(x_N)$. $J_{\pi_k(x_k \ldots x_{\ell-1})}$ depends on π_k only through $\mu_\ell(x_k \ldots x_{\ell-1}X_\ell)$, $\ldots$, $\mu_{N-1}(x_k \ldots x_{\ell-1}X_\ell \ldots X_{N-1})$. This sequence corresponds to the control law π_k after observation of $x_k \ldots x_{\ell-1}$ and is denoted by $\pi_k(x_k \ldots x_{\ell-1})$.

Our goal is to find optimal control laws, that is, control laws that maximise their corresponding gain gamble. In order to do so, we construct a strict partial order on gain gambles, as in (3). This order is derived from conditional lower previsions that describe the uncertain dynamics of the system.

A simple way to describe uncertain dynamics, including learning, is as follows. Suppose at time k we select π_k, and applying π_k up to time ℓ we observe $x_k \ldots x_\ell$. We can now model our knowledge about the state at time $\ell+1$ by a lower prevision on $\mathcal{L}(X_{\ell+1})$, conditional on $x_k \ldots x_\ell$, and depending on the control history $\mu_k(x_k)$, $\ldots$, $\mu_{\ell-1}(x_k \ldots x_{\ell-1})$ and the current control $\mu_\ell(x_k \ldots x_\ell)$. The lower previsions may depend on the full system history, and not only on the current control and state as is the case with Markov decision processes. This allows us to adapt our model according to observations of the system history, and hence, to incorporate learning the system dynamics.

As in Sect. 3 we separate those variables Θ which are not influenced by the control law. Hence, we describe the dynamics by a lower prevision $\underline{P}$ on $\mathcal{L}(\Theta)$, and conditional lower previsions $\underline{P}_{\pi_k}(\cdot|x_k \ldots x_\ell\theta)$ on $\mathcal{L}(X_{\ell+1})$, for each $\pi_k \in \Pi_k$, each $k \leq \ell < N$, each state sequence $x_k \ldots x_\ell$ and each value of θ.The conditional lower previsions are allowed to depend on the control law π_k, but the parameters θ are assumed not to be influenced by the control law.

The separation of act independent variables may appear to be merely a technical assumption. But from Theorem 1 it follows that this separation is *essential* for the principle of optimality to hold. Not separating those variables we recover the weaker order of [3] which violates the principle of optimality.

How to identify act independent variables? Looking at the example invoking the imprecise Dirichlet model for learning dynamics at the end of Sect. 5, these variables naturally arise as the hyper-parameters of the model because they only model prior information. Thus in general, the parameters which are used to represent prior information are act independent.

The conditional lower previsions $\underline{P}_{\pi_k}(\cdot|x_k \ldots x_\ell\theta)$ combine to

$$\underline{E}_{\pi_k}(\cdot|x_k \ldots x_\ell\theta) = \underline{P}_{\pi_k}(\cdot|x_k \ldots x_\ell\theta) \circ \underline{P}_{\pi_k}(\cdot|x_k \ldots x_\ell X_{\ell+1}\theta) \circ \cdots \\ \cdots \circ \underline{P}_{\pi_k}(\cdot|x_k \ldots x_\ell X_{\ell+1} \ldots X_{N-1}\theta) \quad (5)$$

on $\mathcal{L}(X_{\ell+1}, \ldots, X_N)$, as in (2). We can now use (3) to compare control laws after observation of a state sequence. Of course, after such observation it only makes sense to compare control laws with the same control history. Let $\Pi_k(x_k \ldots x_\ell, u_k \ldots u_{\ell-1})$ denote the set of those elements π_k of Π_k for which

$$\pi_k(x_k) = u_k, \quad \pi_k(x_k x_{k+1}) = u_{k+1}, \quad \ldots, \quad \pi_{\ell-1}(x_k \ldots x_{\ell-1}) = u_{\ell-1}. \quad (6)$$

It is convenient to identify $\Pi_k(x_k)$ with Π_k.

Definition 2. *Let π_k, $\rho_k \in \Pi_k(x_k \dots x_\ell, u_k \dots u_{\ell-1})$. We say that π_k is preferred to ρ_k after observation of state sequence $x_k \dots x_\ell$ and application of control sequence $u_k \dots u_{\ell-1}$, and we write $\pi_k >_{x_k \dots x_\ell, u_k \dots u_{\ell-1}} \rho_k$, if*

$$\underline{P}(\underline{E}_{\pi_k}(J_{\pi_k(x_k \dots x_\ell)} | x_k \dots x_\ell \Theta) - \overline{E}_{\rho_k}(J_{\rho_k(x_k \dots x_\ell)} | x_k \dots x_\ell \Theta)) > 0. \quad (7)$$

Using Definition 1, we may select as optimal the set of those control laws which are maximal with respect to the partial order (7).

Definition 3. *A control law $\pi_k \in \Pi_k$ is said to be* optimal *if it is maximal in $\Pi_k(x_k)$ with respect to $>_{x_k}$ for each each $x_k \in \mathcal{X}$. Let $k \leq \ell < N - 1$. The control law π_k is said to be* optimal from time ℓ *if it is maximal in $\Pi_k(x_k \dots x_\ell, \mu_k(x_k) \dots \mu_{\ell-1}(x_k \dots x_{\ell-1}))$ with respect to the partial order $>_{x_k \dots x_\ell, \mu_k(x_k) \dots \mu_{\ell-1}(x_k \dots x_{\ell-1})}$ for each state sequence $x_k \dots x_\ell$.*

5 The Principle of Optimality

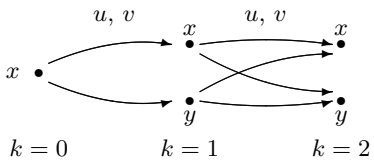

Fig. 1. A simple sequential decision process

Consider the sequential decision process depicted in Fig. 1. At each time k we can choose between actions u and v. We make no assumption on the connection between actions and dynamics. Consider the control law π_0 applying v at time 0, and u if $x_1 = x$ and v if $x_1 = y$ at time 1: $\mu_0(x) = v$, $\mu_1(xx) = u$ and $\mu_1(xy) = v$. The principle of optimality stipulates that if π_0 belongs to the set of optimal control laws from time 0, then the control law $\pi_0(x)$ applying u if $x_1 = x$ and v if $x_1 = y$ at time 1, should belong to the set of optimal control laws from time 1. As a consequence, we can significantly reduce the number of laws we need to consider. To see how this works, assume for instance that ρ_1, specified by $\nu_1(x) = v$ and $\nu_1(y) = u$, is not optimal from time 1. By the principle of optimality, σ_0 and σ_0' specified by $\kappa_0(x) = u$, $\kappa_0'(x) = v$, $\kappa_1(xx) = \kappa_1'(xx) = \nu_1(x) = v$, and $\kappa_1(xy) = \kappa_1'(xy) = \nu_1(y) = u$, cannot be optimal from time 0, because otherwise ρ_1 would be optimal. Hence, knowing optimal control laws from time $\ell + 1$ can help reducing the search space when looking for optimal control laws from time ℓ. Of course, we can do this only if reducing the search space does not change the set of optimal elements we end up with: our notion of optimality must be *insensitive to omission of non-optimal elements*. Fortunately, the insensitivity property holds: it suffices that every non-optimal element is dominated by an optimal element [1].

Theorem 1 (Principle of Optimality). *Let $k < N$ and $\pi_k \in \Pi_k$. For any $k \leq \ell < N$, it holds that if π_k is optimal from time ℓ then it is optimal from time $\ell + 1$, whenever all of the following conditions are satisfied:*

- *The conditional lower previsions $P_{\pi_k}(\cdot|x_k \dots x_\ell\theta)$ are linear, for all $k \leq \ell < N$, all values of θ, and all state sequences $x_k \dots x_\ell$.*
- *There is a $T \subseteq \Theta$ such that $\underline{P}(f(\Theta)) = \inf_{\theta\in T} f(\theta)$ for any gamble $f(\Theta)$.*
- *For any $x_{\ell+1} \in \mathcal{X}$ it holds that*

$$\underline{P}(P_{\pi_k}(I_{X_{\ell+1}=x_{\ell+1}}|x_k \dots x_\ell\Theta)) > 0. \tag{8}$$

In short, the principle of optimality holds if the imprecision is concentrated on the act independent variable Θ, and if it is of the following type: θ is only known to belong to some set $T \subseteq \Theta$. Thus, whenever the imprecise model is described by a *set* of precise models $\{P_{\pi_k}(\cdot|x_k \dots x_\ell\theta)\colon \theta \in T\}$ *and these precise models are connected through a conditioning variable* θ, the principle of optimality applies when using the preference order (7). Imprecise probability models are often expressed as a set of precise models. The theorem says that we should look for an act independent variable which parametrises this set.

This situation obtains exactly when we use an imprecise Dirichlet model in order to represent learning: the conditional linear previsions are

$$P_{\pi_k}(f|x_k \dots x_\ell\theta) = \textstyle\sum_{x_{\ell+1}\in\mathcal{X}} f(x_{\ell+1}) \frac{s\theta^{\mu_\ell(x_k\dots x_\ell)}_{x_\ell x_{\ell+1}} + n^{\mu_\ell(x_k\dots x_\ell)}_{x_\ell x_{\ell+1}}(x_k\dots x_\ell,\pi_k)}{s+N^{\mu_\ell(x_k\dots x_\ell)}_{x_\ell}(x_k\dots x_\ell,\pi_k)} \tag{9}$$

for any gamble f on $X_{\ell+1}$, and the unconditional lower prevision is

$$\underline{P}(g) = \inf\{g(\theta)\colon \theta^u_{xy} > \epsilon, \textstyle\sum_{y\in\mathcal{X}} \theta^u_{xy} = 1\} \tag{10}$$

for all gambles g on Θ. Let's briefly explain these expressions.

$n^u_{xy}(x_k \dots x_\ell, \pi_k)$ denotes the number of transitions from state x to y by applying u, in the sequence $x_k \dots x_\ell$ subject to π_k, and $N^u_x(x_k \dots x_\ell, \pi_k) = \sum_{y\in\mathcal{X}} n^u_{xy}(x_k \dots x_\ell, \pi_k)$. Equation (9) is the predictive lower prevision on $X_{\ell+1}$ which arises from an independent product of precise Dirichlet models on the transition probabilities from state x_ℓ applying $\mu_\ell(x_k \dots x_\ell)$ after having observed $x_k \dots x_\ell$ subject to control law π_k [4]. Observation of transitions from one state do not influence our knowledge about transitions from another state. This motivates the use of an independent product of Dirichlet models, each model modelling transitions from a particular state.

The variable $s > 0$ determines the adaptivity (lower s means faster learning), and $\theta^u_{xy} > 0$, $\sum_{y\in\mathcal{X}} \theta^u_{xy} = 1$, determine the prior transition probabilities from state x to y applying control u. Equation (10) says we only know *a priori* that the lower probability of any transition is at least $\epsilon > 0$ (such that (8) holds). S and Θ determine prior information about the dynamics and are obviously not influenced by π_k: they are act independent.

6 Discussion and Conclusions

The principle of optimality (together with the insensitivity property) yields an efficient recursive algorithm, dynamical programming, that calculates optimal control laws, reducing the global optimisation problem which requires a search over all Π_k, to a sequence of local optimisation problems requiring only a search over $\mathcal{U}$. In this way we achieve an exponential speedup in determining the set of optimal control laws. Unfortunately, we have to omit an exact description due to lack of space. Its construction is well described in [1].

Since control laws depend on the full system history, the algorithm runs over all possible histories, and not simply over all possible states as in the case without learning. As a result, we still need an exponential time (but even so, the search space remains exponentially smaller than Π_k). This is inevitable also in the classical approach, even considering sufficient statistics. Thus, a direct implementation is only feasible for small systems. On the other hand, in our learning approach precision increases with time. Hence, with longer time horizon the incomparability of control laws will be less likely, and the size of the optimal set will tend to stabilise, unlike the methods proposed in [3, 1].

Acknowledgements

This paper presents research results of project G.0139.01 of the Fund for Scientific Research, Flanders (Belgium), and of the Belgian Programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office. The scientific responsibility rests with its author.

References

1. Gert de Cooman and Matthias C. M. Troffaes. Dynamic programming for deterministic discrete-time systems with uncertain gain. Submitted to the International Journal of Approximate Reasoning, December 2003.
2. Robert Givan, Sonia Leach, and Thomas Dean. Bounded-parameter Markov decision processes. *Artificial Intelligence*, 122:71–109, 2000.
3. David Harmanec. Generalizing Markov decision processes to imprecise probabilities. *Journal of Statistical Planning and Inference*, 105(1):199–213, June 2002.
4. J. J. Martin. *Bayesian Decision Theory and Markov Chains*. John Wiley & Sons, New York, 1967.
5. Jay K. Satia and Jr. Roy E. Lave. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.
6. Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
7. Chelsea C. White and Hany K. Eldeib. Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749, 1994.